Who Wants a Piece of Me?
Reconstructing a User Profile from Personal Web Activity Logs

Salman Elahi, Mathieu d'Aquin and Enrico Motta

Knowledge Media Institute, The Open University
Walton Hall, Milton Keynes, MK7 6AA, UK
{s.elahi, m.daquin, e.motta}@open.ac.uk

Abstract. One of the major contributions of Web 2.0 to our lives is adding the ease of information sharing and socializing online. Users communicate and exchange with many different websites, resulting in large amounts of personal data being transferred, as many "identity fragments" spread over multiple services. This fragmentation in personal information exchange makes it difficult for users to really comprehend, and therefore control, their own data. An envisaged solution to this issue is the deployment of "user-centric" profile management systems, where a global profile is being managed by the user, and granted access to for various services. However, it is unclear whether, in the open, complex environment of the Web, where data exchange can take many different forms, such an approach would be applicable. In this paper, we present an experiment using real, large-scale personal Web activity data to reconstruct a global user profile from the many fragments of personal information exchanged over a period of time. While this demonstrates the potential value of "user-centric" profile and identity management, it also helps us identifying new research issues and challenges that will be faced by such an approach.

1   Introduction

In our daily lives, we interact with different types of people, e.g., colleagues, friends, family and strangers as well. Usually we manage this sort of interactions well, as each interaction has its own space and restrictions. However, when it comes to online interactions, all exchanges happen in one, single, open space, with known, trusted websites having the same status as unknown ones, and data exchange constantly occurring, at a pace impossible for us to fully comprehend. We refer to this phenomenon as the *fragmentation of personal data exchange*: where many different "destinations" of the data receive various fragments of personal information over time, without a global management and understanding of the data. This is currently the way data exchange happens on the Web, with obvious consequences on data control, ownership and, of course, privacy.

User centric profile management systems deem to be a solution to this issue [1]. Current research initiatives in this area consider an approach where, instead of fragments being sent to various destinations, a global profile for the user is

maintained in the system and managed by the user himself. Parts of this profile can be requested by various websites, with the user being given the possibility to control these accesses and to keep track of them directly within the profile management system. However this idea has not been tested in a realistic environment, corresponding to actual interactions happening on the Web. Users send a huge amount of personal information in small fragments during interactions with different websites, through online forms, tweets, Facebook comments, automatically gathered information, etc. There is a need to check whether the idea of a global profile management system could be applicable in such a complex and heterogeneous environment, and what are the challenges to be faced to actually realize and deploy such a system. It is very important to stress upon the fact that we use the term 'profile' here in the sense of a collection of personal data, rather than of a the set of user interests and preferences as used for example in [2].

In this paper, we present results from an experiment relying on the actual Web activity logs collected from a user over a period of time. The idea here was to collect all the fragments of personal information sent by this user during several weeks, and to try to reconstruct from these fragments a coherent global profile. Besides requiring the design of a number of tools, the main contributions of this experiment are, on the one hand, to demonstrate the feasibility of personal information exchange relying on the assumption that one, global user profile is being maintained by the system, and on the other hand, to identify the issues and research challenges that we will need to face to realize such a system.

2   Related Work

The idea of global user profiles goes back to late 1970s, when generic user modelling servers and applications were looked at as potential solution to personalisation issues [3]. However, they still lack wide acceptance due to the complexity involved in adapting them from one organisation to another. Identity management shares a thin boundary with the aspects of personal information management we are looking at. The most popular approach for "user-centric" identity management is OpenID[1]. OpenID is a protocol that provides a unique ID to each user to be used over various services, a given ID being authenticated by a chosen identity provider. It therefore supports users in reducing the distribution of their identity data. OAuth[2] coupled with OpenID provides secure exchange of data without disclosing users' credential to third party websites.

However, while some attributes of personal information can be passed through OpenID, OpenID and OAuth are essentially concerned with the problem of authentication and leave the management of personal information exchange to the third party websites being accessed. Going a step further, there are a few initiatives to realize user-centric identity management frameworks, including Windows CardSpace[3], LiberryAlliance[1], Higgins I-Card[2], etc. These frameworks provide central

---

[1] http://www.openid.net/
[2] http://www.oauth.net/
[3] http://www.microsoft.com/windows/products/winfamily/cardspace/default.mspx

places for storing and managing personal information (i.e., profiles), to which external websites are granted access. In this sense, they comply with our notion of profile providers. [4, 5] concern flexible user profiles in mash up environments. They discuss frameworks to provide decentralised and domain independent user models which can be used across applications. Morpho [4] facilitates the interoperability of user profiles among various applications, i.e., eliciting a user profile from one application and transforming it to be useable by another application. GUMF (Grapple user Model Framework) [5] is focused on e-learning environments for profile interoperability. [6, 7] also discuss very similar approaches to [4, 5]. They consider a framework called SUPER (Semantic User Profile Management Framework) to aggregate the user information spread across multiple data silos to model semantic user profiles. SUPER is focused on the retail domain. While they still suffer from a number of limitations (e.g., in iCard based frameworks, personal data is still fragmented, in the sense that profile information is "boxed" according to the websites requesting it), a critical issue for us is that these frameworks are only initial development models, which have not been tested in real-life, open Web scenarios.

FOAF+SSL [8] is an authentication protocol to provide secure exchange of distributed information. It uses the SSL layer in modern day Web browsers to create a secure conduit of information between two parties. Through the use of Semantic Web technologies (FOAF), FOAF+SSL goes a step further in terms of user-centrality: it allows a user to create his own profile and host it himself for applications to access securely. However, while very promising, FOAF+SSL is also in an initial phase of development and has not been widely adopted yet. In this sense, the assumptions on which it relies have not been validated on realistic cases.

Here, we realize an experiment to re-construct a global user profile from the fragments of personal information sent by a real user to a large variety of websites, in order to test the feasibility of the approaches mentioned above. A similar approach of constructing a user profile from different fragments is presented in [9], which re-creates global-user profiles from profile information available on the Web. In contrast, we focus here on re-constructing a user profile from the fragments of personal information exchange present in the user's Web activity log. The way to collect such an activity log is described in the next section.

3   Monitoring Web Activity for Data Exchange

These days, with Web 2.0 applications, information is constantly being exchanged between users and websites. However, mechanisms to monitor these information exchanges are not commonplace, or generally very limited (i.e., browsers' Web history). It is indeed a difficult task currently to keep track of what pieces of information a user shared knowingly (i.e., through direct interaction with websites) or unknowingly (e.g., through syndication and push-based client applications, or as an implicit side effect of explicit Web activity). As mentioned above, our goal is to collect actual data on personal information exchange on the Web to experiment with.

---

[1] http://www.projectliberty.org/liberty/about/
[2] http://www.eclipse.org/higgins/

To get such fragments of personal information we have developed a tool to monitor a user's personal activities on the Web. Technically, it takes the form of a Web proxy system which, installed on the computer of the user, intercepts and records any communication occurring with the "external web" through the HTTP protocol. The corresponding logs are then encoded in an RDF format, relying on purpose-built "HTTP Ontology" (see Figure 1).
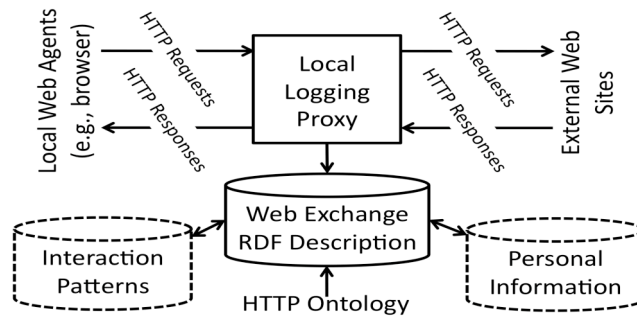

Fig. 1. Overview of the Web activity monitoring system.

We ran this programme on the computer of the second author of this paper for 2.5 months, which yielded 9GB of data corresponding to more than 100 Million RDF triples; representing over 3 Million HTTP requests and the corresponding responses. To make sense and extract meaningful information out of these logs, the data is automatically annotated, stored and manipulated using semantic technologies (RDF, OWL). Taking advantage of ontological reasoning it is possible to classify this data into different categories, e.g., agents, hosts, requests, websites, etc. As a result, we have a structured version of this complex data ready to be used for the profile re-construction phase of our experiment. This rich amount of data also helped us to analyse different aspects of a user's online activities and data exchange [10].

4   Re-constructing a User Profile from Heterogeneous Fragments

This section describes the experiment we conducted to test the hypothesis of a global semantic user profile using a bottom-up approach, from the data collected using the method described above.

4.1   Basic Model

Web activity log data can be considered as three interlinked sets of data: the set of hosts (website main URLs), the set of attributes (request parameters in URLs and

POST requests), and sets of values (values from the user for the request parameters). Each triple *<host, attribute, value>* represents a unique entry in our Web activity log. There can be more than one instance of a host, attribute pair, which however can uniquely be identified through the linked value (i.e., triples are unique, see Figure 2).
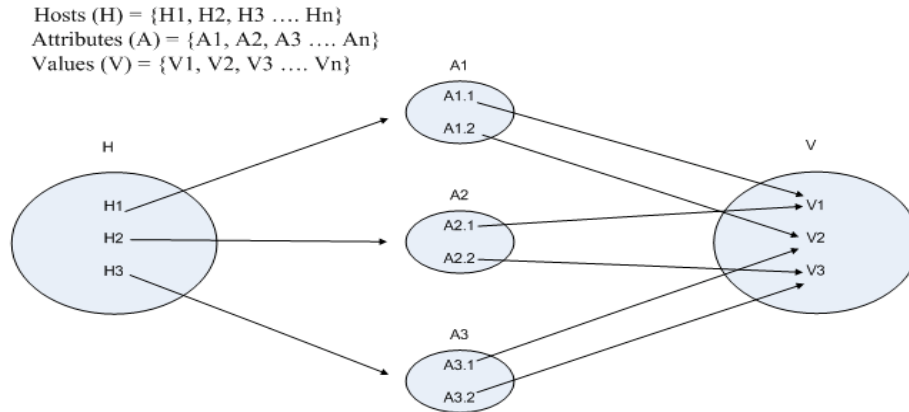

Fig. 2. The Basic Model: data from the user sent to hosts using attribute-value pairs.

This relationship can be further explained with the help of the following examples from the log (Table 1).

Host	Attribute	Value
www.google.com	http://www.google.com/search#q	yahoo+sandbox
www.google.com	http://www.google.com/search#q	yahoo+searchmonkey
www.google.com	http://www.google.com/search#q	jisc+access+identity
ees.elsevier.com	http://ees.elsevier.com/ijhc/update.asp #username	md'quin-992
ees.elsevier.com	http://ees.elsevier.com/ijhc/update.asp #email	m.daquin@open.ac.uk

Table. 1. Web activity triples.

Here we can see that a host (*www.google.com*, *ees.elsevier.com*) receives various values through various attributes (*q*, *username*, *email*), with the same attribute possibly being used multiple times, and the same value possibly being sent through different attributes, and to different hosts.

Our experiment consists in re-constructing a profile for the user from which the activity log data originates. We use a simple representation for this profile, made of attribute-value pairs (e.g., *firstName=Mathieu*). Of course, each attribute can be associated with multiple values, and a value can be used in multiple attributes. We start with an empty profile, with no attribute defined.

In order to populate the profile with attributes and values for these attributes, we introduce the notion of mapping from the profile to the activity log. A mapping defines a relationship between an attribute in the profile and a set of attributes used in the Web activity triples defined above. The existence of a mapping indicates, first that the mapped profile attribute should be present in the profile, and second, that any value associated with the corresponding "data attribute" for a given host in the activity log should be used to populate the profile for the mapped profile attribute. These mappings can be one-to-one or one-to-many from profile attributes to data attributes, as shown in Figure 3.
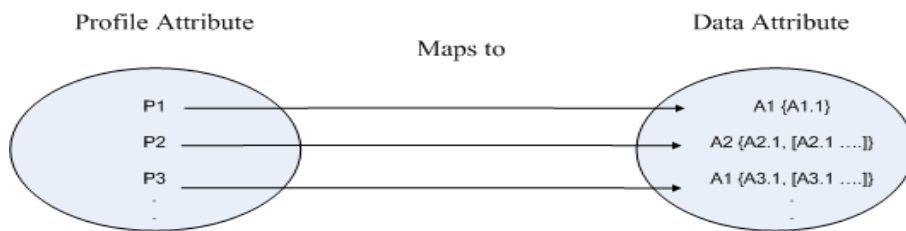

Fig. 3. Mapping profile attributes to (log) data attributes.

For example, if we want to aggregate all the fragments of information about the user's login IDs for different websites as a profile attribute called *username*, we create a mapping between the *username* profile attribute (which might or might not exist before) and selected attribute(s) used by these different websites to collect usernames, as shown in Figure 4.
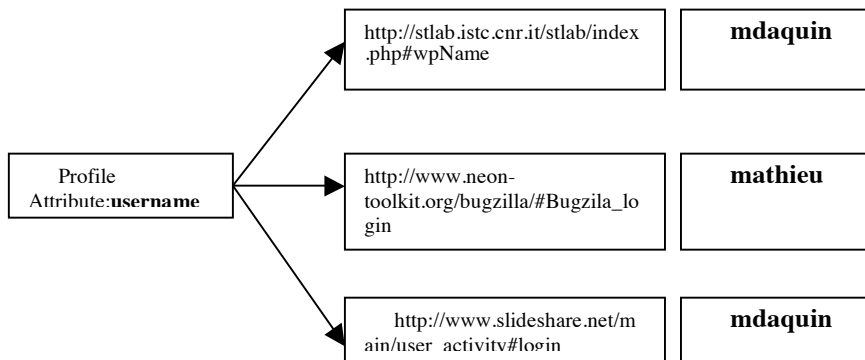

Fig. 4. Example of profile to data attribute mapping for *username*.

Here we can see that the profile attribute *username* maps onto three different data attributes having two different values: *mdaquin* and *mathieu*.

4.2   An Interactive Interface to Create Profile to Web Activity Log Mappings
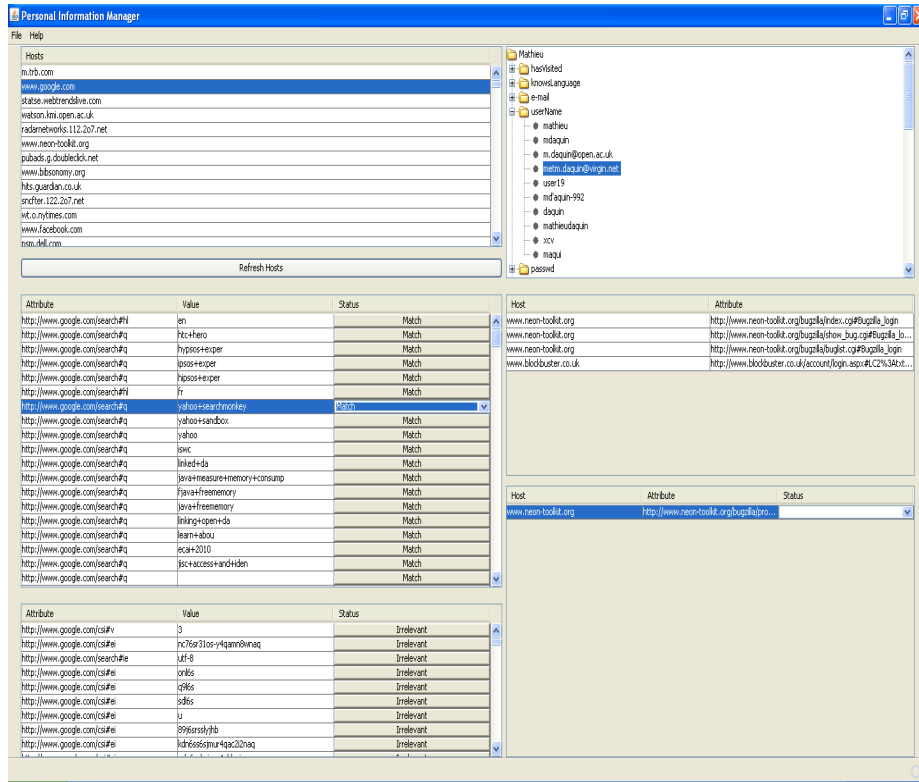

Fig. 5. Profile generation tool using mappings between the user profile attributes (top-right corner) and Web activity log attributes (bottom-left corner). The tool lists all the relevant hosts in the top-left corner and also provides suggestions of mappings based on already entered values in the bottom-right corner.

We have developed a tool to help the user to re-construct a profile based on fragments of personal information present in the Web activity log, using the notion of mappings described above. This tool (see screenshot in Figure 5) helps the user to explore the activity log data in a convenient way, providing an ordered list of hosts, to which information about data (attributes and values) received is attached. The user can then categorise each triple into one of three categories: *Unknown* (no mapping defined yet), *Match* (at least one mapping exist), and *Irrelevant* (to be discarded as not relating to personal information. Usually technical parameters used by websites fall into this category). Initially, each triple is placed in the *Unknown* category, and hosts are ranked according to the number of attribute-value pairs they have in the *Unknown* and *Match* category.

Categorizing an attribute-value pair in a given host as *Match* automatically triggers a dialogue to create a new mapping. The user can then chose an existing profile attribute to be part of the mapping, or create a new one by simply giving its

name. Once the mapping created, the tool will automatically 1- create the profile attribute if it does not exist, 2- populate this profile attribute with any value attached to the attribute of the selected Web activity log triple and 3- update the categorization of dependent triples and the order of hosts.

Through this simple mechanism, it is possible for the user to re-create his own profile, by simply selecting elements of the Web activity log and indicating to which part of the profile they relate. In addition, as we keep track of the created mappings, it is always possible to explore the data further by inspecting to which hosts a particular value of the profile has been sent, and what other information this particular host has received.

Moreover, in order to make this process easier for the user, the tool also generates suggestions of mappings, by identifying in the Web activity log the attributes that have received values already known in the profile. In this way, as soon as, for example, the e-mail address of the user has been mapped once, the tool will list all the other host-attribute pairs in the data which seem to correspond to the e-mail address, as they have received the same value.

4.3 Results

We had 33,098 *<host, attribute, value>* triples as part of our initial Web activity log. These triples were extracted from the 3 Million requests mentioned in Section 3 by discarding all irrelevant information, regarding for example the HTTP request headers, requests without data exchange, as well as duplicate requests. On an informal tone, the user found the process of re-constructing his own profile intuitive, involving and informative. For example, it was found that around 50% of the hosts that had received personal information were completely unknown to the user (i.e., he never explicitly visited them). A large proportion of them are Web marketing and traffic analysis sites such as *www.google-analytic.com*, which topped the list of the hosts having received the largest volume of information from the user.

As can be seen in the Figure 6, the generated profile encompasses a large variety of aspects of the user's personal information, ranging from relatively insignificant pieces of information (e.g., screen resolution), to more critical data (e.g., username, e-mail address, phone number). Indeed, within a relatively short time (2-3 hours) the user managed to create 36 profile attributes mapped onto 1,108 data attributes out of the 33,098 triples in our log. This has been made easier through the mapping suggestion feature of the tool described in the previous section. Of course, in a real life environment, the time and effort necessary to create these mapping should be even further reduced; using sophisticated semi-automatic mapping creation and sharing mechanisms.

Despite a reasonably short adoption period, the mapping process and profile re-construction appeared straightforward to the user. In particular, we could not find any relevant piece of data in the Web activity log that could not fit in our basic model of profile and mappings. However, it is obvious that this experiment takes a number of restrictive assumptions on the way user profiles are represented and managed. This experiment also helped us identifying in which way such assumptions would be

challenged in a realistic profile management system, and therefore, what are the research issues to be tackled towards the deployment of such a system.
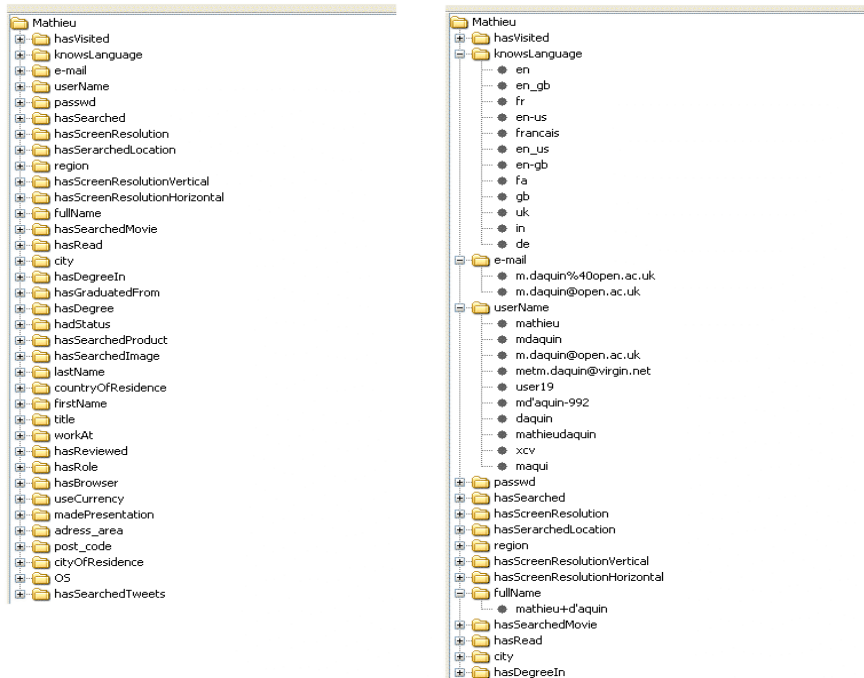

Fig. 6. Collapsed profile (left), Expanded profile (right).

5   Discussion

The results from our experiment are encouraging for what concerns the possibility of applying profile management systems in an open Web environment. Indeed, we can see from the previous section that, even with a relatively simple model of profiles and mappings, building a global view on the user's personal information, as exchanged with various websites, is feasible in reasonable time and without major inconsistencies appearing.

However, a major goal of this experiment was also to better understand the issues that profile management systems would face in a realistic Web environment. Indeed, our model relies on simplistic assumptions concerning the way profiles are represented and managed. In the course of running the experiment, we identified a number of high-level issues where these assumptions appeared too restrictive. We discuss here these research issues and challenges.

5.1 Profile Representation

In this paper, we have used a very simple representation for a user profile: attribute-value pairs. While this appears very basic and unsophisticated compared to some models of user profiles, our experiment indicates that a large portion of Web users' personal data can be represented using this simple model. However, there are a number of cases where a less straightforward model would be more satisfactory.

The most obvious of these cases are the ones where substructures need to be represented in the profile, i.e., where attributes should be attached to other objects than the user. For example, elements concerning some journal articles that the user has reviewed appear in the data. In such cases, we create an attribute *hasReviewed* and populate it, e.g., with the identifier of the paper in the target system. However, the data also includes additional information about the article itself, such as its title and authors. Such information cannot obviously be attached to the user, but should be represented, for example, through triples of the form *<Mathieu, hasReviewed, paperX>*; *<paperX, hasTitle, "Great Paper on Cool Things">*; *<paperX, hasAuthor, "Don T. Exist">*. Even more, in this example, a ternary relation would be necessary to fully represent this piece of information and encode the fact that *Mathieu has reviewed paper X with score Y*.

While a model able to cope with such information would not be difficult to create using modern semantic technologies, the main issue here comes from the mappings with the requests from external websites. Indeed, as we use a bottom-up approach to re-construct a profile from the mappings with real data, a semantic profile management system would need mappings to fulfill information requests from websites, and answer them using data from the user profile. Our simple profile model is very easy to map to the currently simple model of parameters used to exchange data in HTTP. However, making the representation of the profile more sophisticated will inevitably lead to an increase in the complexity of the mappings, making it more difficult to create them and to maintain them.

At a higher level, other representational issues have emerged from running our experiment. In particular, it appears that some websites have obtained different values from the same attributes over time. In many cases, this does not mean that the profile of the user should contain several values, but that the value changed, and so that the underlying implicit profile evolved. Representing this dynamicity of the user profile (i.e., having temporal profile) is critical in a profile management system, as some websites might refer to previous "versions" of a profile, or request information about the history of a given value (i.e., the user addresses in the last 5 years).

Another high level issue concerns the inherent property of a profile as being "multi-faceted". Indeed, it appears clearly that the reason why the *e-mail* attribute has multiple values is that these values relate to several "sub-identities" of the user (e.g., personal vs. professional). The goal of a profile management system is to avoid fragmentation with the use of a global user profile. It is however important for such a system to acknowledge this multiplicity, by somehow making this global profile "context dependent".

5.2 Enriching the User Profile with Multiple Information Sources

Above, it is assumed that only two sources of information contribute to the user profile: the user himself, and the corresponding websites through mappings. However, there exist a number of additional resources that could be used to enrich the profile with additional information and, as a consequence, help in making it easier to manage.

First, the user might want to be given better possibilities to edit the profile himself, modifying values and structuring the information in a way that fits his own model and view on his personal data. While this is not fundamentally a difficult problem, the issue is, like in the case of structured representations above, in the mappings. Indeed, websites might have requested the edited information before and so, mappings might exist between the data in its current form and the way these external websites would consume it. Maintaining the consistency of these mappings by automatically propagating the applied changes in this context represents a challenge.

Second, information about the user profile might be already openly available on the Web for the system to collect and use. Some information might be exposed through social networking websites, or other personal websites of the user. One interesting perspective here is to consider how much the current availability of some pieces of personal data can be used as a basis in deciding on granting access to it to some particular websites. In addition, information is not only available about the profile itself, but also about elements of the profiles and about the potentially requesting websites. It would be easy to realize, using a data source such as *geonames.org* for example, that the city the user lives in (Milton Keynes) is part of the country UK. Many other connections of the sort could be realized concerning the user's employer, address, relations, etc. relying on the many sources of information available as linked data. These connections would appear crucial in a profile management system willing to help the user in understanding the possible consequences of disclosing a particular piece information to a given website, showing him what inferences this website could draw from such information, considering the data it is expected to have access to already.

6 Conclusions and Future Work

In this paper we have presented our findings from an experiment on re-constructing a user profile from the fragments of data he sent to various websites as part of his normal Web activity. The goal of this experiment was to consider the feasibility and applicability of a "user-centric" profile management system relying on a global user profile aggregating all of such fragments. We developed several tools to realize this task and conducted the experiment on the basis of user data collected during a period of 2.5 months. From this data, we managed to re-create what would be the implicit global user profile for this user, showing how it could be mapped to the way websites currently request this data. While this shows the potential of the considered profile management systems, it also helped us identifying high-level research issues that such

systems should face as this approach has already been used in [11] to help the user with trust and privacy issues during online data exchange.

Further steps in this research include considering the functional aspects of a user-centric profile management system for the Web, using a simulated environment where interactions with websites could be replayed "as if" they were mediated by this system, thus giving us an insight on the necessary protocols to be put in place.

References

1. Renato Iannella. *Social Web Profiles*. Position paper at the 1st international workshop on social network interoperability, SNI 2009.

2. Roman Y. Shtykh and Qun Jin. *Dynamically constructing user profiles with similarity-based online incremental clustering*. International Journal of Advanced Intelligence Paradigms, 377-397, 2009.

3. Alfred Kobsa. *Generic User Modeling Systems*. User Modeling and User-Adapted Interaction, 2004.

4. Erwin Leonardi, Geert-Jan Houben, Kees van der Sluijs, Jan Hidders, Eelco Herder, Fabian Abel, Daniel Krause and Dominikus Heckmann. *User Profile Elicitation and Conversion in a Mashup Environment*. First International Workshop on Lightweight Integration on the Web, ComposableWeb'09.

5. Fabian Abel, Dominikus Heckmann, Eelco Herder, Jan Hidders, Geert-Jan Houben, Daniel Krause, Erwin Leonardi and Kees van der Slujis. *A Framework for Flexible User Profile Mashups*. International Workshop on Adaptation and Personalization for Web 2.0, AP-WEB 2.0 2009.

6. Riddhiman Ghosh and Mohamed Dekhil. *Mashups for semantic user profiles*. Poster, International World Wide Web Conference, WWW 2008.

7. Riddhiman Ghosh and Mohamed Dekhil. *I, Me and My Phone: Identity and Personalization Using Mobile Devices*. Technical Report, Digital Printing and Imaging Laboratory, HP Laboratories Palo Alto, HPL-2007-184, 2007.

8. Henry Story, Bruno Harbulot, Ian Jacobi and Mike Jones. *FOAF+SSL: RESTful Authentication for the Social Web*. 1st Workshop on Trust and Privacy on the Social and Semantic Web, SPOT 2009.

9. Matthew Rowe. *Interlinking Distributed Social Graphs*. Linked Data on the Web Workshop, LDOW 2009.

10. Mathieu d'Aquin, Salman Elahi and Enrico Motta. *Personal Monitoring of Web Information Exchange: Towards Web Lifelogging*. Web Science Conference, 2010.

11. Mathieu d'Aquin, Salman Elahi and Enrico Motta. *Semantic Monitoring of Personal Web Activity to Support the Management of Trust and Privacy*. ESWC 2010.